FEBS Letters 586 (2012) 1942–1952

journal homepage: www.FEBSLetters.org

Review

BMP signaling in wing development: A critical perspective on quantitative image analysis

Alexander Brooks [a,1], Wei Dou [b,1], Xiaoying Yang [b], Tara Brosnan [a], Michael Pargett [c], Laurel A. Raftery [a,*], David M. Umulis [b,c,*]

[a] School of Life Sciences, University of Nevada, Las Vegas, 4505 S. Maryland Parkway, Las Vegas, NV 89154-4004, USA
[b] Agricultural and Biological Engineering, Purdue University, 225 S. University St. West Lafayette, IN 47907, USA
[c] Weldon School of Biomedical Engineering, Purdue University, 206 S. Martin Jischke Drive, West Lafayette, IN 47907, USA

ARTICLE INFO

Article history:
Received 13 February 2012
Revised 23 March 2012
Accepted 24 March 2012
Available online 31 March 2012

Edited by Joan Massagué and Wilhelm Just

Keywords:
Quantitative image analysis
Mathematical modeling
Wing development
Drosophila
Normalization

ABSTRACT

Bone Morphogenetic Proteins (BMPs) are critical for pattern formation in many animals. In numerous tissues, BMPs become distributed in spatially non-uniform profiles. The gradients of signaling activity can be detected by a number of biological assays involving fluorescence microscopy. Quantitative analyses of BMP gradients are powerful tools to investigate the regulation of BMP signaling pathways during development. These approaches rely heavily on images as spatial representations of BMP activity levels, using them to infer signaling distributions that inform on regulatory mechanisms. In this perspective, we discuss current imaging assays and normalization methods used to quantify BMP activity profiles with a focus on the Drosophila wing primordium. We find that normalization tends to lower the number of samples required to establish statistical significance between profiles in controls and experiments, but the increased resolvability comes with a cost. Each normalization strategy makes implicit assumptions about the biology that impacts our interpretation of the data. We examine the tradeoffs for normalizing versus not normalizing, and discuss their impacts on experimental design and the interpretation of resultant data.

1. Introduction

Bone Morphogenetic Proteins (BMPs) regulate development and homeostasis of numerous tissues in both vertebrates and invertebrates (Wu & Hill, 2009). The model organism *Drosophila melanogaster* is an outstanding system to investigate novel mechanisms for BMP regulation, due to the sophisticated tools available to test for BMP signaling in whole tissues. BMPs regulate many events throughout the *Drosophila* life cycle, including embryonic axial patterning (reviewed in [1,2]), growth and patterning of appendage primordia (reviewed in [3,4]), and maintenance of germline stem cells (reviewed in [5,6]), to name a few. However, current debate about the biological robustness of BMP-mediated patterning in *Drosophila* tissues is fueled, in part, by differences in data interpretation, which can arise from different approaches to quantitative comparisons of developmental patterns between control and test samples [7–14].

Recently, developmental analyses have been applied to determine the biophysics of transport [12,15], the cell biological response to extracellular BMPs (e.g. [16]), and to build mathematical models of tissue morphogenesis [7,8,10,12–18]. Such computational studies rely upon the quantification of microscopic images obtained as assay results (for a general perspective on bioimage informatics, see [19]). Both intrinsic biological variability and extrinsic experimental variability are present in these images, creating challenges that need to be overcome in order to draw conclusions with adequate statistical power. However, few investigators have addressed basic questions about how the quantification should be done. Here we offer our perspective on quantitative assays for spatial distributions of BMP activity, using the specific case of *Drosophila* wing development, and suggest how the results apply to other developmental contexts.

In all systems studied, the spatial and temporal distribution of BMP activity is tightly controlled for proper development and homeostasis. Regulatory interactions occur at all levels of BMP signaling, including ligand activity and availability, ligand–receptor binding, and the lifetime of activated signal transducers [2]. For example, multiple extracellular proteins (e.g. Crossveinless-2, Short Gastrulation/Chordin, Noggin, and Collagen), bind competitively to secreted BMP ligands and restrict both ligand diffusion and local binding to signaling receptors. By their very nature, these regulatory

* Corresponding authors. Addresses: School of Life Sciences, University of Nevada, Las Vegas, 4505 S. Maryland Parkway, Las Vegas, NV 89154-4004, USA (L.A. Raftery), Agricultural and Biological Engineering, Purdue University, 225 S. University St. West Lafayette, IN 47907, USA (D.M. Umulis).
E-mail addresses: laurel.raftery@unlv.edu (L.A. Raftery), dumulis@purdue.edu (D.M. Umulis).
[1] These authors contributed equally to this work.

0014-5793/$36.00
http://dx.doi.org/10.1016/j.febslet.2012.03.050

interactions alter the spatial and temporal distribution of BMP activity across a tissue. The dynamic complexity of this tissue-wide network has led to computational tests of molecular mechanisms. Although we are in early stages of this exciting approach, such studies open the door to understanding BMP signaling in the systems level context of tissue function integrated in space and time.

Computational tests of mechanistic models rely on quantitative data collected from biological experiments. To interrogate the dynamics of BMP signaling across whole tissues, immunofluorescent staining and microscopic imaging are the principal tools. Both input (extracellular BMP ligands) and output (phospho-RSmad or target gene expression) can be quantified using established biological assays, e.g. immunostaining or transgenic, fluorescent proteins [8,12,15,9,20,21]. Increasing numbers of studies of tissue patterning by BMPs and other morphogens use numerical representations of images for a quantitative comparison between control and experimental populations [8–10,12,22–28]. However, individual laboratories follow their own methods for extracting quantitative data from microscopic images, often without explicitly describing them in resultant publications. Manipulation of quantitative data has similarly proceeded by ad hoc approaches.

It is widely recognized that a biological assay transforms the true biological information in the tissue, such as the total number or concentration of a molecular species, into a quantitatively distinct observation such as fluorescence intensity patterns in a microscopic image. In a typical study, control and experimentally manipulated animals are reared in parallel, tissue specimens are isolated and processed in parallel, and digital images are obtained using a sophisticated fluorescence detection system, such as a laser-scanning confocal microscope (LSCM) [9,10,12,21]. For each population multiple specimens are imaged, and these image data sets are compared. To augment such comparisons, population statistics are used to test for significant differences between control and experimental populations [29].

Recently, some investigators have critically evaluated whether common image quantification methods are appropriate for the analysis of image data from experimental manipulations of morphogens acting in *Drosophila* embryonic patterning [10,23,30–32]. In Section 2, we use these evaluations as a starting point to focus specifically on the challenges that emerge when examining BMP activity distributions. The selection of assay carries with it different sources of biological and assay variability that critically impact decisions about the design and interpretation of experimental results.

In particular, the observed data has added variability from material, observational and conceptual error [33]. These errors impact experiment-to-experiment variability, sample-to-sample variability, and specimen variability, in which an experiment is composed of aggregated samples from manipulated and control populations, and each sample population is composed of multiple specimens. The cumulative variability impacts our ability to discriminate any real differences between the experimental and control specimen populations. These mixed sources of variability lead to measurement error that hinders analysis of biological pattern formation by BMPs.

Increasing sample population size through normalization of data sets might improve an investigator's ability to quantify spatial patterns. However, normalization can alter the data in unexpected ways and potentially bias interpretation of the biology. In Section 3, we take a theoretical view of the kinds of variation that are present in the data, and the post-processing approaches that are taken to minimize variation within and between populations of data. Herein we aim to offer our perspective by focusing on the following questions: (1) How many specimens are needed for satisfactory discriminatory power between control and experimental populations? (2) Should data be normalized, and if so, which type of image normalization provides the most faithful representation of the biological information? (3) What properties of a morphogen gradient should be measured and used in comparison tests? This discussion provides a starting point for future consideration of methods to quantitate morphogen gradients and evaluate BMP signaling during development.

2. Tradeoffs in assay design: examples from *Drosophila* wing primordia

While tradeoffs exist in any quantitative imaging experiment, we focus on patterning of the Drosophila wing primordium. Flies are holometabolous insects, in which massive growth occurs during the larval stages, and then adult structures develop their final form during a non-feeding pupal stage [34]. Adult appendage development begins in discrete cell populations that form structures called imaginal disks [35]. Imaginal disks grow extensively during larval stages, with spatial patterning events that restrict regions of the tissue to specific adult fates. Each wing develops from a specific region of a wing imaginal disk, the wing primordium (Fig. 1). By unknown mechanisms, imaginal disk growth is coordinated throughout the organism, so that each appendage is appropriately sized for the overall size of the individual fly. Furthermore, patterning and growth are coordinated, so that all aspects of adult structure are proportional, or scaled to size [36]. Recent computational studies of BMP signaling in wing development have focused on how the distribution of BMP activity scales with tissue size [7,9].

Proteins of the *Drosophila* BMP signaling network are highly conserved with mammalian BMP signaling proteins, but are generally named via an idiosyncratic genetic nomenclature. The BMP2/4 ligand ortholog is Decapentaplegic (Dpp); for wing primordia, the BMP 5/6/7/8 class ligand is Glass Bottom Boat (Gbb) (recently reviewed in [37–39]). Just as in mammalian BMP signaling, ligand binding recruits a heteromeric complex of a type II receptor, which is a constitutive serine–threonine kinase, with a type I receptor that is a conditional serine–threonine kinase [40]. The transient ligand-type I–type II complex permits type II receptor-mediated activation of the type I receptor, which then phosphorylates a BMP R-Smad. Phosphorylated R-Smad accumulates in the nucleus, oligomerizes with the co-Smad, and interacts with other transcription factors to directly regulate expression of numerous genes. The *Drosophila* BMP R-Smad ortholog is Mothers Against Dpp (Mad). Antibodies that detect the phosphorylated form of Mad (pMad), are commonly used to assay the levels and spatial distribution of BMP activity in fly tissues, although some analyses use transgenic proteins tagged with fluorescent proteins.

2.1. Impact of assay variability on quantitative image data

To frame the discussion on tradeoffs between approaches, it is helpful to consider the types of variability that leads to error in a typical biological assay. There are two primary types of error in an assay: random error of measurement and systematic error of measurement. Allchin et al. further categorize experimentally-introduced errors according to their origins: material error, observational error, and conceptual error [33].

Material error originates from physical conditions that cannot be perfectly controlled. The alterations might be multiplicative and/or additive to the true biological data, scaling the molecular concentration data. Multiplicative error poses a significant problem since each subsequent step in the process of imaging and quantification has the potential to multiply earlier errors, leading to compounding errors akin to compounding interest for a bank account.

Observational error originates from human involvement in the process of research and refers to inadvertent errors based on the

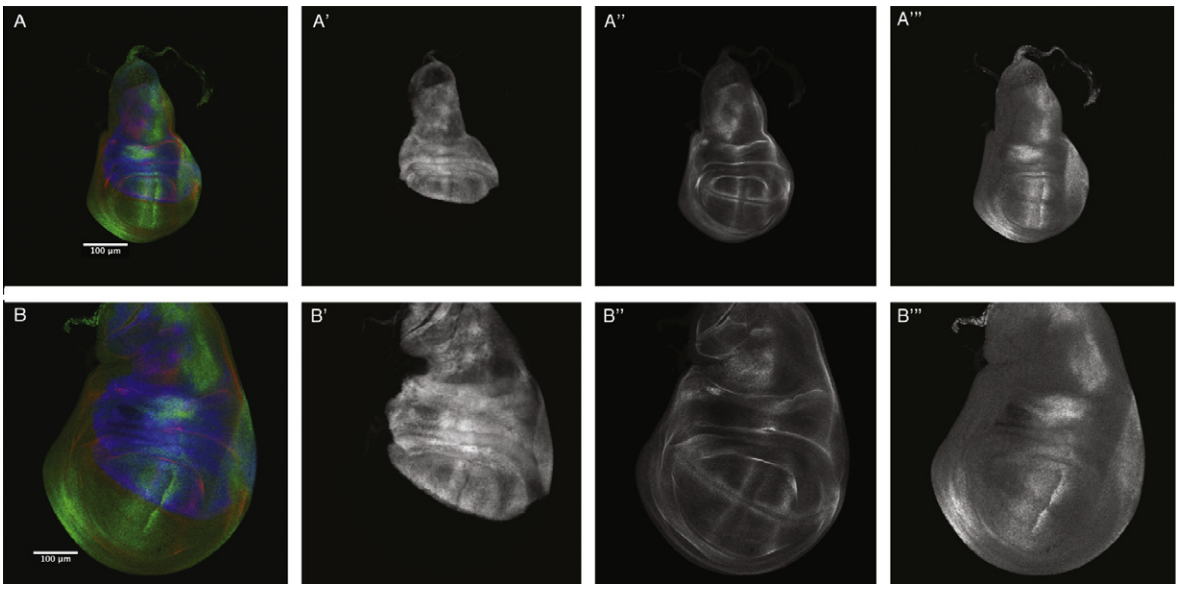

Fig. 1. Natural size variation in wing imaginal discs. Row (A) and row (B) show fluorescence images of two wing discs isolated from the same cohort of larvae (genotype: y w; Ap-Gal4, UAS-GFP/+; brk-lacZ/+) and processed in the same trial. (A and B), merged images. (A′ and B′): GFP reporter (blue in merged images) for dorsal cells. (A″ and B″) Delta (red in merged images), indicating the pattern used to stage tissues (mouse α-delta, DHSB, #C594.9B; Cy5-goat α mouse, Jackson Laboratories, #115–175-003). (A‴ and B‴) PhosphoMad (green in merged images, rabbit α -phospho-Smad1/5, Ser463/465, Cell Signaling, #9516, Alexa 568-goat α-rabbit, Invitrogen, #A11011). Raw images from each channel acquired sequentially with a Nikon A1R LSCM at 20× magnification (0.75 NA) 512 × 512 pixel resolution, and 8-bit pixel depth. Merges and maximum intensity projections produced with Nikon elements software. Image adjustments were made simultaneously to all panels with Photoshop.

observer interaction with data acquisition. For example, the specific parameter settings used for image acquisition (image metadata, [41]) can alter the signal-to-noise ratio, resulting in observational error in the raw images obtained.

Conceptual error arises predominately through steps that are taken post-acquisition and depends on the soundness of theoretical formulation for data manipulations. An example is whether images are scaled prior to data analysis (see Section 2.3.1). Further conceptual error can result from choices of statistical methods, precision of computational routines and choice of normalization strategy.

2.2. Selection of signaling assays

The selection of assay has significant impact on the type and level of experimental error, which ultimately impacts the type of questions that can be addressed by the data and the level of statistical trustworthiness of the data. For example, live imaging of a Dpp-fluorescent protein fusion eliminates many steps needed for a typical antibody staining, but the introduction of the fusion protein into an organism could alter the state of the system leading to results that do not reflect the endogenous situation.

In general, fluorescence microscope imaging is the method of choice for quantitative image analysis, but it has numerous pitfalls, which are extensively reviewed elsewhere [42–44]. We focus on the experimental design that provides a fluorescent marker for BMP activity.

2.2.1. Direct detection of fluorescent BMP protein distribution

Fluorescent protein (FP)-tagged proteins are attractive reagents to detect BMP distribution in living samples [15]. Direct FP fluorescence provides a measure of the number of molecules present at the location of each pixel [45,46]. However, a FP tag has significant potential to interfere with protein function because of its size; for example, a GFP tag is a 238 amino acid moiety. Additional deviation from native biological morphogen distributions may arise from the methods used to express the tagged protein within the tissue. For example GFP::Dpp transgenes are expressed in wing primordia using a binary, transgenic expression system (Gal4-UAS, [47]) that has apparently spatially appropriate expression conferred by a portion of the *dpp* cis-regulatory region; but with this system, the fusion protein does not fully rescue *dpp* mutant phenotypes. Thus, a transgenic construct may provide an inaccurate assay via a system that is related to, but not equivalent to, the endogenous situation, even though material error is reduced because the immunostaining reactions are removed from the assay.

2.2.2. Indirect detection of BMP activity

Because additional regulation occurs at the levels of ligand accessibility to receptor, ligand–receptor complex lifetime, and receptor binding to R-Smad, assays of intracellular responses are often used to evaluate the distribution of BMP signal *activity*. The most widely used assay is immunodetection of pMad (Fig. 1, example profiles extracted in Fig. 2). Several antibody reagents are available, and all give qualitatively similar results [48–50]. However, this assay is not universally appropriate, because it is an indirect detection assay that only works on fixed tissues.

Expression of BMP target genes, or transgenic reporters derived from them, is another common assay. Transgenic reporters expressing FPs, such as *dad-RFP* [12], can be used to follow the dynamics of BMP activity distribution in living specimens [12]. However, transcriptional responses are the integrated output from interactions of multiple transcription factors (e.g. regulation of *vg* by BMP and Wg pathways [54]). Thus, reporter or target gene expression may not be faithful to the distribution of BMP activity.

2.3. Assay contributions to data variability

2.3.1. Size variation

A widespread challenge to whole tissue imaging is size variability within a population of organisms (examples in Fig. 1). Currently, investigators choose between three approaches to manage size variation. The first is to perform quantitative comparisons without consideration of the overall size of each specimen [51]. A second approach is to modify culture methods to strictly limit

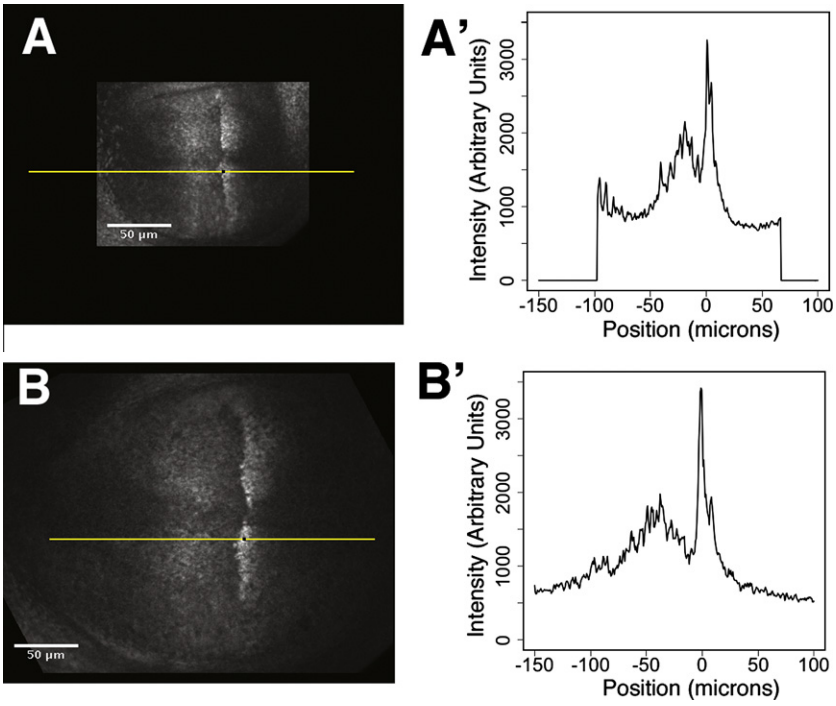

Fig. 2. BMP activity profiles. (A), (B) PhosphoMad staining in wing primordium regions from specimens in Fig. 1. Yellow stripes indicates where the profiles in (A′), (B′) were quantified. (A′), (B′) Phospho-Mad intensity profiles from a ventral position15% offset from the dorsal/ventral boundary. Images acquired at 40× magnification (NA 1.3), and images processed as for Fig. 1. Each value in the profile is the mean of a 1 × 8 pixel window. The images were annotated in Fiji. The profiles were generated using the statistics software R.

nutritional variation that may alter tissue size [9], which has the advantage of not perturbing the image data, but limits the numbers of specimens that can be collected for each sample.

The third approach to size variability is to convert all image data to the same scale, either during image processing or while processing the extracted quantitative data. Data from wing primordia may be scaled using a length metric that is moderately independent of BMP activity (length parallel to the anterior–posterior boundary). This treatment systematically transforms the data from each specimen, but has not been assessed for fidelity of the processed data to the original data. For this reason, scaling image data during initial processing may not be appropriate.

2.3.2. Day to day variability in assay parameters

Variability in assay parameters is a fundamental feature of experimental science. Often cell biologists measure assay variability with replicate assays on the same biological material [29]. For imaging experiments, each assay repeat requires new specimens, so that replication is not possible. Because a new sample of specimens is obtained each time an imaging assay is repeated, we will use "trial" to refer to each parallel treatment of samples.

A common approach to manage assay variability is to restrict comparisons to control and experimental specimen populations that were assayed in the same trial. For tissues acquired by microdissection, this approach places a strong constraint on either the number of specimens for each test population (e.g. [9]), or the number of populations directly compared (e.g. [21]). Additional control specimens may be included to measure background levels from non-specific staining [8,12]. Alternatively, background levels may be minimized for each specimen with offset adjustments during image acquisition [41], which confounds quantitative comparisons.

Data normalization to manage assay variability is increasingly used for quantitative analyses of image data. This approach facilitates the accumulation of larger specimen populations from multiple assay trials, and thus provides a two-pronged approach to limiting variability. Data can be normalized in the form of an image, or as extracted quantitative representations. Different methods of normalization are available; in Section 3 we consider how simulated biological data is transformed by normalization and how this may affect interpretation of the data. Specifically, Section 3 demonstrates how some normalization methods may alter the population characteristics, such as the mean.

3. Quantifying BMP activity distributions

To detect differences between control and experimental populations or to constrain mathematical models, knowledge about the contributions of the assay to final variability can greatly inform early decisions on how images should be processed. The types, estimated amounts, and sources of variability all impact our experimental design and also need to be considered in the selection of image comparison and analysis metrics. We start by considering local variability within an image.

3.1. Impact of local biological variability on quantitative image data

In addition to specimen-to-specimen variability within a sample, local variation occurs within a single specimen. Within each cell, reaction rates fluctuate over time as a result of dynamic variations in local molecular concentrations and variability in local extrinsic factors such as the behaviors of neighboring cells. Such molecular scale events affect ligand production rates, uptake rates, transport rates, and other processes that occur at the cellular scale [8]. Additional pixel-to-pixel variability is contributed by the fluorescent imaging modality used for image acquisition [52]. Thus, both biological and assay variability may appear as fluctuations in the intensity profile from a single specimen (Fig. 3), which otherwise would be expected to appear smooth. These fluctuations

create observable differences in measured fluorescent intensity at the same relative location. A widespread and simple approach to manage such internal variability is to generate each point in the intensity profile by averaging the values for a group of adjacent pixels (e.g. Fig. 2). This method reduces point-to-point variation within an image, with the assumption that molecular levels are locally uniform.

Despite the inherent variation from cell to cell, each specimen exhibits a profile pattern that is stereotypical for a genetically related population. Important insights into the underlying molecular

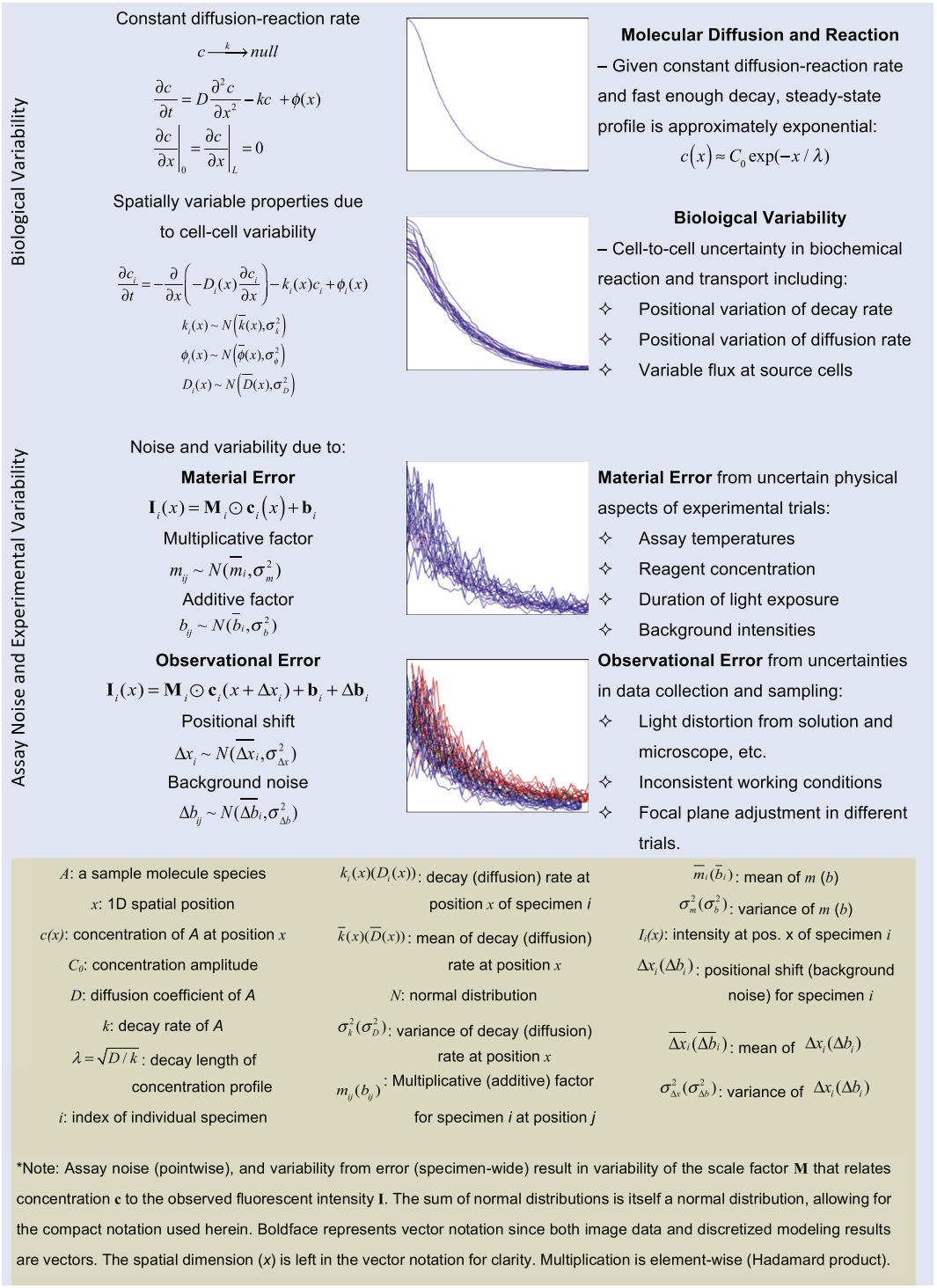

Fig. 3. Sources, mathematical descriptions and summary of error contributions to the measurement of BMP signaling. Top: An example protein distribution in a tissue, such as the wing primordium, is shown here to demonstrate how biological patterns have inherent variability and how variability accumulates during experimental assays. For all plots, the horizontal axis represents spatial position along the anterior/posterior tissue axis and the vertical axis is the concentration (biology) or measured fluorescent intensity (assay). A total of 20 simulated samples is shown in the plot, where the blue lines are individual simulated samples and red is the mean profile. The following assumptions are applied in the system: molecule A is (1) present at the left-most region of the sample in constant amount, (2) diffuses across the X axis at rate D, and (3) decays with rate k.

events are gained from detection of pattern alterations exhibited throughout an experimentally manipulated population. Thus, in whole tissue studies, a common goal is to extract overall pattern information from the inherently variable data. A specimen's profile pattern can be smoothed using a mathematical spline function [12,49], but it is unknown how this affects fidelity to the original data. We focus on the effects of data normalization on the profile pattern. In Section 3.2, we computationally simulated a molecular concentration pattern, in order to compare the efficiency of specific normalization approaches to processing quantitative image data.

3.2. Evaluation of quantitative image analysis methods

Whether the goal is to detect phenotypic differences or to estimate the uncertainty for measured biophysical parameters, one must consider how many samples are needed to establish statistical power for the conclusions. We use this concept of statistical power to frame our analysis of normalization approaches, for the simple reason that most biological experiments have limited sample size. We analyze a simulated data set generated using a reaction–diffusion model for Dpp movement across posterior cells of wing primordia (Fig. 3), based on earlier work by Bollenbach et al. [8].

3.2.1. Model data sets used for evaluation

To create the simulated dataset, the reaction–diffusion model includes sources of biological variability that are time-invariant to represent cell-to-cell differences in the primordium that confer local variability in the biophysical parameters for production ϕ, diffusion D, and kinetic uptake/decay k (Fig. 3) [8]. The results of this model calculation create a baseline "biological" concentration profile that we refer to as the Ground Truth (GT) data. Multiplicative and additive errors are added to the "specimen" as a whole and pointwise, to approximate error introduced by each step of the assay, the image acquisition, and the processing steps used to extract quantitative intensity profiles. The output of this second model, simulating the addition of assay material and observational error, produces a population of morphogen distributions that each represents the "observed" (OB) data equivalent to a profile from an LSCM image of specimen fluorescence.

This approach gives us two simulated data sets: (1) the GT data that contains only biological variability, and (2) the OB data that contains both biological and experimental variability. These data are useful as a test case, but we recognize that our data simulation includes assumptions that contribute to conceptual error, and that machine error (truncation error in computer decimals) leads to minor material error in the numerical solutions to the model. Estimation of minimum population size and evaluation of normalization methods are entwined and need to be evaluated together since the estimate for minimum population size depends on the normalization strategy.

3.2.2. Overview of three methods for data normalization

Whether or not image data should be normalized is controversial, and rightly so. This data processing step is potentially dangerous, depending on how the process transforms the original data. While each normalization method generates profiles that superficially resemble the GT data (Fig. 4A–E), each method makes different implicit assumptions about the underlying biology that directly impact the resultant interpretation.

A common normalization method, which we call the "anchor point" (AP) method, is to set the maximum intensity value (from a single pixel or a local average of pixel intensities) from each specimen equal to the same maximum value, often using arbitrary units. This method imposes the *implicit* assumption that each individual in the sampled population can exquisitely control the level of morphogen or signal at a spatial location (method AP1). The assumption is not generally valid. A related strategy, used to investigate Bcd patterning, is to force the maxima and minima (or the average of subpopulations of data points) to be equal to one and zero respectively, implying two biological assumptions about the data (method AP2) [26,31,32]. Lastly, an extension of the AP method is to set subpopulations of the image data to an average of the top 5% and bottom 5% of the pixels and this is equivalent to a max and min filter (AP3). AP3 type normalization may also occur during image acquisition from the scaling and gain algorithms in LSCM controller software.

The "integral" (IN) normalization method divides each profile by the value of the background-subtracted morphogen profile's integral over the domain of measurement [8]. The resulting profiles all have equal integrals, which imposes the implicit assumption on the normalized data that each specimen contains the same number of the molecular species assayed. This assumption has not been tested in a biological context, and is not consistent with our current understanding of stochastic, physical variability between cells, as discussed in Section 3.1.

"Model-based" (MB) normalization imposes the implicit assumption that specimen to specimen variation is primarily due to assay-dependent errors in measurement for each observation, e.g. variations in overall fluorescent intensity for an immunostained sample. This approach assumes that the true biological data is transformed into the observed data, by a mathematical function for the biological assay. The parameters for the function will vary significantly from trial to trial and to a lesser extent between specimens within a trial, and may contain both multiplicative and additive errors. The overall goal of this normalization strategy is to minimize the error between the observed intensity data and the expected result from the "model" prediction. For a linear model, this produces a new data set that minimizes the squared error between individual specimens and the mean for the population of specimens and has been referred to as a variance minimization approach or a χ^2 minimizer. MB normalization that used a linear model was first used to investigate the reliability of Bicoid (Bcd)-mediated embryonic patterning and subsequently to investigate dorsal surface patterning by BMPs [10,22,31].

MB normalization requires an estimate "model" or "function" for the assay parameters that transform the biological data into observed data. Then, computational methods seek the optimal parameters for the function that minimizes the difference between each observation and the model estimate for the observation. To establish the approach, Gregor et al. showed that immunostaining for Bcd and live-imaging of embryos with a Bcd--GFP fusion protein produce fluorescent intensity distributions that are linear functions of concentration [22]. A linear relationship between fluorescent intensity and concentration gives the function $I_n = A_n * c_n + B_n$, where I_n is the intensity, c_n is the concentration for each image n, and A_n and B_n are the scaling parameters that transform the concentration data into the measured fluorescent intensity. Each A_n and B_n can be different for each image within a population due to intrinsic and extrinsic variability.

If one assumes that the majority of image-to-image systematic variability is experimentally introduced, then the observations I_n can be explained by variability in the parameters for the assay (A_n, B_n) that transform the true biological mean $c_{mean}(\boldsymbol{x})$ into each observed fluorescent intensity profile. Following the minimization routine (Fig. 4), the result of the normalization is the selection of each A_n and B_n, an estimate for each $c_n(\boldsymbol{x})$ consistent with the method, and $c_{mean}(\boldsymbol{x})$, the mean concentration for the data set. The variability in the intensity is absorbed by the model parameters that scale the true mean biological data (an unknown, but calculated quantity) to the model predicted intensity. While the method is theoretically based on inverting the function that transforms true biological data (unknown) into the observable quantity, it is

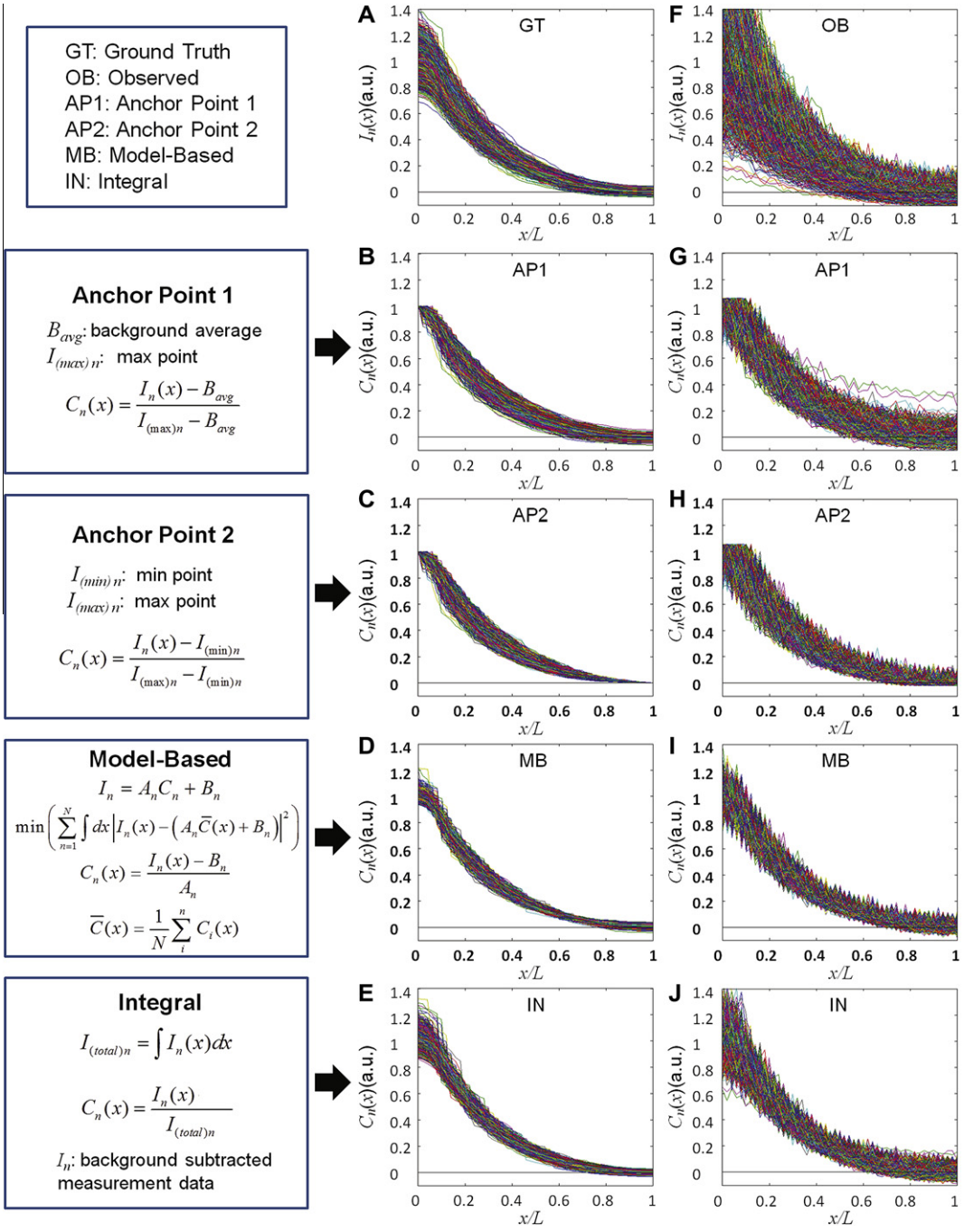

Fig. 4. Comparison of normalization methods. Left column: summary of approach. (A–E) Normalization of GT data by AP1, AP2, IN, and MB approaches. (F–J) Same as (A–E) except applied to OB data.

incapable of distinguishing between sources of variability. To summarize, the implicit assumption in model-based normalization is that the "true" biological data is robust and highly reproducible between individuals within a population, and that the majority of the observed variability is caused by the assay. Since the method cannot distinguish between sources of individual-to-individual multiplicative variability, it also *imposes* this assumption on the data being normalized and can lead to normalized data with lower estimates for the variability than is present in the specimen population.

3.2.3. Evaluation of normalization strategies

The major goals for normalization are to reduce the variability introduced by the observer, improve confidence in measured quantities, and reach more reliable conclusions about the biology. We used our model data sets (See Section 3.2.1) to determine whether normalization achieves these goals, or instead hinders our ability to measure biological parameters and detect differences between populations. The model data give us the advantage of starting with an initial data set that we know with certainty. We refer to this initial data as GT data. The data with variability added to each point yields the OB data. Normalized data sets are generated by each strategy: OB (not normalized), AP1, AP2, AP3, IN, and MB, and each normalized data set is compared against GT data to estimate the impact for that strategy. Since normalization establishes profiles on a relative (normalized) scale with "arbitrary units (au)", the mean of each population is adjusted so that they all have

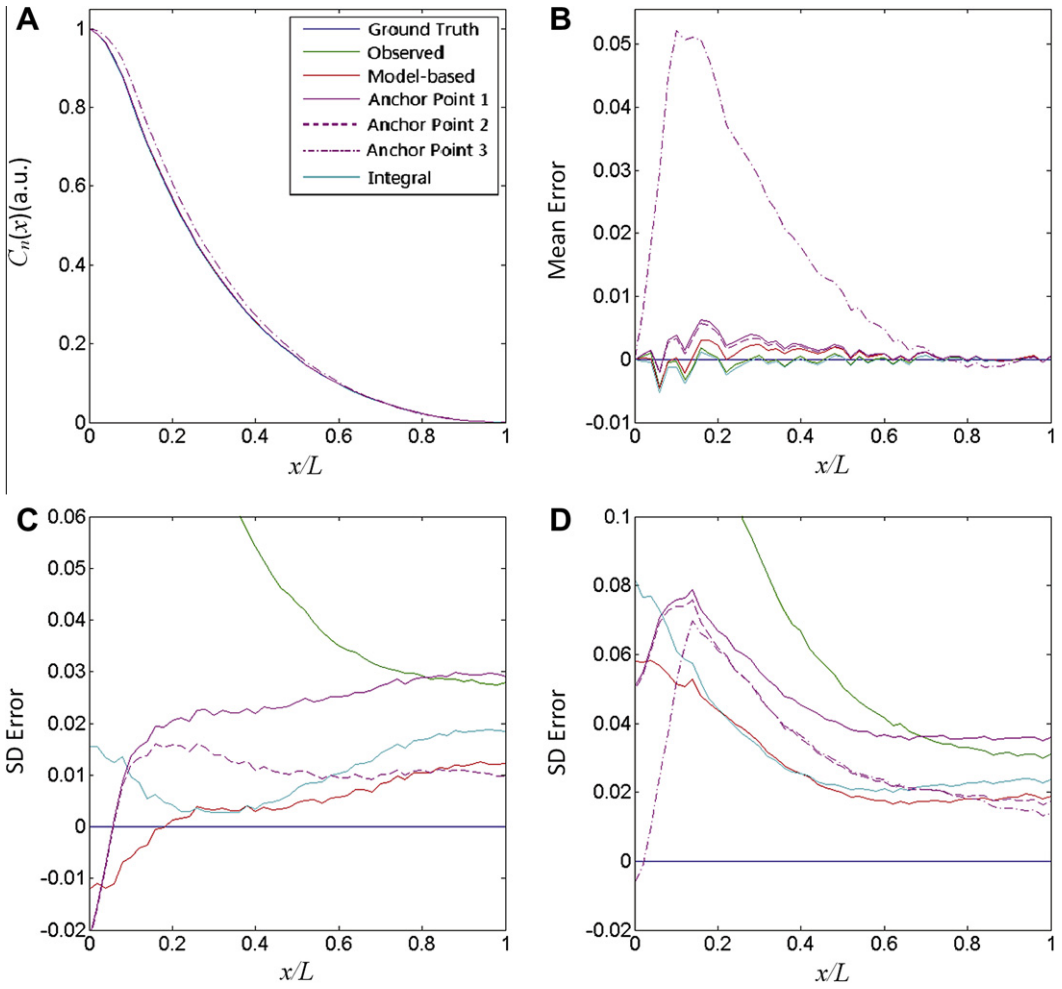

Fig. 5. Normalization impact on mean and standard deviation of sample data. (A) Comparison of mean profiles after normalization. (B) Error in mean. (C) Estimation of standard deviation after normalization for 5% assay noise. (D) same as (C), except with 10% noise.

equivalent scale and can be compared directly to each other and the GT data.

To acquire good estimates for the population mean and standard deviation for each data set, the model was simulated 5000 times. While this is an unrealistic number in any morphogen imaging experiment, simulating large numbers is necessary so we can later estimate minimum populations sizes needed to achieve statistical significance in a typical experiment.

3.2.4. Normalization's potential pitfalls and successes

Normalization allows for larger populations sizes by aggregating data from different trials that may provide greater statistical power to measure variables within and between samples. A primary concern for normalization is that the process can systematically alter the data, and in some cases our interpretation of the biology. To investigate normalization's positive and negative impacts, we measured how each normalization method discussed above changes the population-level statistics for GT and OB data. GT data still contains simulated biological variability that will impact the output of each method in different ways and this test serves as a "worst case" scenario for normalization's impact on interpretation. OB data provides a more realistic trial with data that contains both biological and experimental variability. In Fig. 4A–E, it is clear how each normalization method impacts the GT data and leads to quantitatively different distributions of morphogen patterns. AP1 and AP2 pinch the distributions at the max (AP1) or both max and min (AP2) with higher variability in the profiles between the anchor points. Both MB and IN methods reduce the variability in the GT data and demonstrate how normalization can impact data and our subsequent interpretation of data in the hypothetical case when there is NO experimental variability (the perfect assay). When 5% normally distributed noise (see Fig. 3) is applied pointwise across each individual for imaging, detector noise, stochastic reaction events in the assay, etc. and 20% multiplicative and additive trial-to-trial variability is applied to yield OB data, the normalization methods differentiate relative to each other in the resultant distributions. AP1 and AP2 lead to greater amounts of variability far from the anchor points, whereas MB and IN "appear" closer to the GT data.

We observed qualitatively similar mean morphogen profiles before and after normalization, except for AP3, (not shown in Fig. 4) which overestimates the concentration distribution relative to other methods (see Fig. 5A). To confirm this observation, we plotted the difference between the mean after normalization and the mean of the GT data (Fig. 5B). There was no measurable difference between the mean for the GT data and any of the means from the OB, AP1, AP2, IN, and MB data, even with large population sizes that exceed what is feasible in a typical experiment. The application of multiplicative and additive error to the ground truth data did not lead to a different mean (Fig. 5A,B), and this is expected since we applied normally distributed error, consistent with our expectation for each step of the assay.

We then calculated the error in the standard deviation (σ) in concentration for each data set and plotted the error in σ as a function of

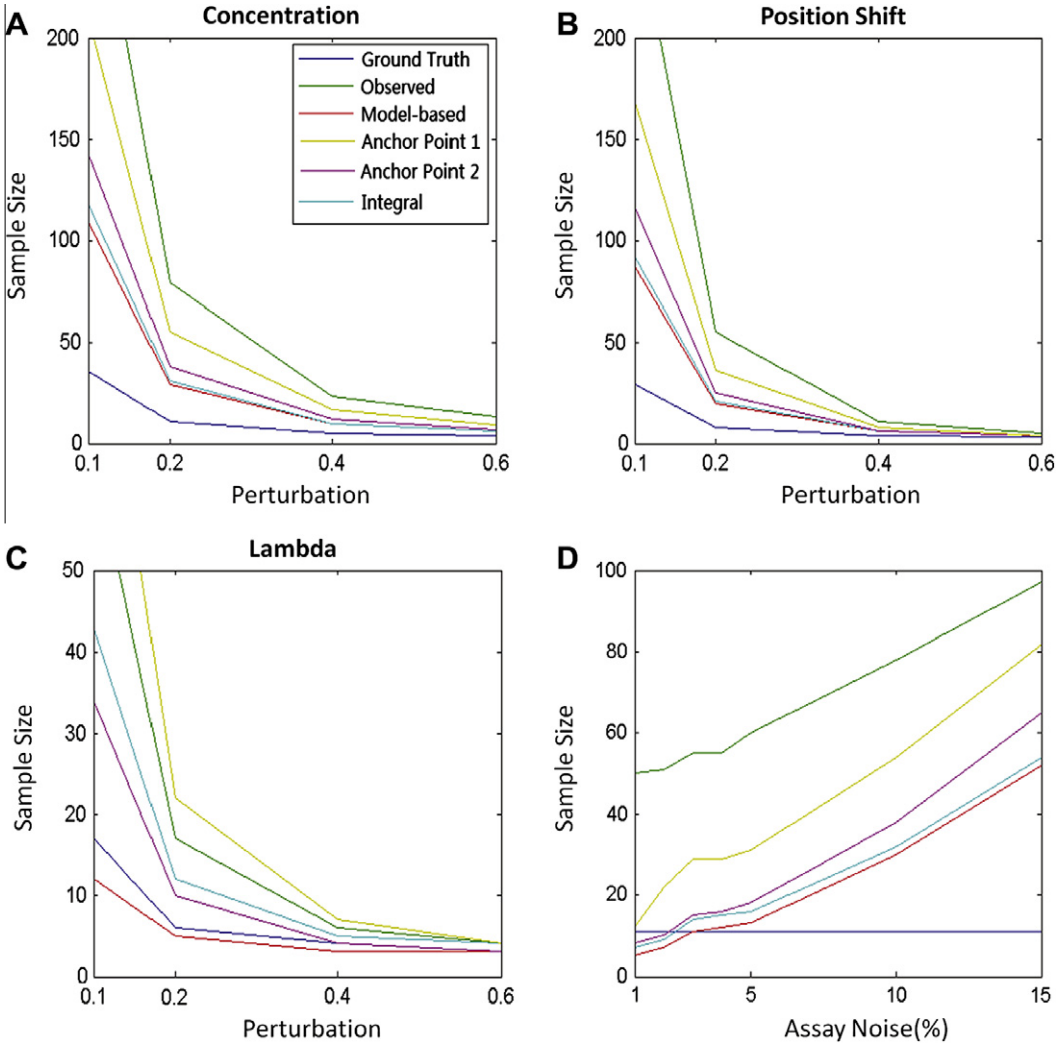

Fig. 6. Sample size estimation. (A–C) Estimated sample sizes for measuring a (A) difference in concentration (data from each population are labeled 10%, 20%, 40%, and 60%, the percent decrease in the mean decay rate *k*), (B) position shift at a threshold, and (C) changes in decay length λ. (D) Estimates for sample sizes to detect concentration differences as a function of the assay variability.

position (Fig. 5C, D). The OB data set leads to the greatest σ at all points tested and showed the greatest difference when compared against the ground truth data. AP1 and AP2 normalization led to varying results in the estimate of σ that depends strongly on relative location. Consistent with the previous evaluation of the method for Bcd data [26,31], anchoring by max and min leads to overestimates of σ between the anchor points. In effect, variability that exists in the max and min are eliminated at the cost of transmitting the variability to intermediate spatial positions. Gregor et al. thoroughly described how this can impact the interpretation of the underlying biology of Bcd patterning [31]. Normalization by the IN and MB methods lead to better approximations for σ, but there are tradeoffs. With 5% assay noise, IN leads to an overestimate of the σ relative to GT throughout the distribution, whereas MB underestimates σ near the peak concentration and overestimates it throughout the remainder of the profile. With 10% noise, all methods lead to standard deviations greater than GT data, and the MB approach overestimates by the smallest amount relative to the other methods.

3.2.5. Normalization's impact on detecting differences between morphogen profiles

The ability to detect differences between test versus control populations is central to experimental design, and normalization can have a large impact on our ability to draw conclusions from the acquired data. In particular, it is essential for studies of morphogen patterning robustness and reproducibility. Unfortunately this field is peppered with different approaches to assessing the extent of similarity between the extracted patterns from control and test populations. Normalization impacts our ability to detect differences between extracted patterns, so we extended the method to estimate the minimum numbers of samples needed in each population to have a power of 0.9 and $\alpha = 0.05$ as a function of the strength of the biological perturbation and the three profile parameters used: intensity, shift in spatial position at a threshold, or lambda, the decay length.

We simulated the biophysical morphogen model using variable values of the decay/endocytosis constant that alters the range of morphogen movement (on average). The values are similar to recent experimental results that suggest Pentagone impacts the range of BMP activity in the wing primordia [53]. We simulated the data from experimental populations where the decay rate is reduced from wild-type levels. The estimate for the number of samples needed to detect a difference between control and experiment are shown for each normalization method and GT data in Fig. 6 A. Small differences in the decay rate require relatively large populations to reject the null hypothesis for all

methods, with the greatest sample size needed for the unnormalized OB.

The selection of profile metrics also influenced the sample size needed. If either a shift in position or lambda was used for comparison between control and experiment, the minimum population size was lower than when detecting differences in concentration (Fig. 6B,C). The minimum requirement for lambda varied from population sizes of 3 to >30 depending on the extent of the perturbation. Measures in the spatial shift of the gradient at a threshold level required intermediate population sizes. Unexpectedly, fewer samples were needed to detect a difference after normalization by the MB method than were needed for the GT data. This suggests that normalization can go too far and eliminate some natural biological variability.

Lastly, we examined how the results for sample size changes as a function of the assay error caused by noise. Additionally, since normalization can impact the distributions of GT data, there is a point where normalization should not be used. To test this, we estimated sample sizes to detect a difference in concentration for assay noise (pointwise variability) that increased from 1% up to 15%. For all runs, the trial-to-trial variability was fixed at 20%. For assays with very low noise, normalization reduces σ, leading to estimates that are lower than the minimum required for GT data. At 5% noise and above, normalization reduces the required sample sizes relative to OB data.

In summary, these simulations show that normalization is useful to reduce the sample size needed in order to detect differences between control and experimental populations of sample images. If the ability to obtain samples is a limiting factor in experimental design, AP1, and AP2 are the least desirable methods, as they require the most samples to detect the same level of difference between control and test samples in an experiment. The most desirable method, in our opinion, is the MB approach that provided consistent estimates for mean and standard deviation throughout the exponential gradient that were close to the GT data. Two profile metrics, λ and positional shift at a threshold concentration, require fewer samples to detect differences in patterns than measuring differences in concentration.

4. Summary and perspective

The primary goal for this evaluation is to initiate an open dialog on the quantification of BMP signaling during development. With the growing recognition of the intricate networks for BMP signaling, there will be even greater reliance on quantitative imaging technologies and analysis methods to interpret the underlying biology and elucidate the regulatory networks. We emphasize that assay design for image quantification should be considered at the outset. Our evaluation using a simulated data set suggests that some normalization methods are more appropriate than others, and that experimental design should include consideration of how the normalization might affect the resultant data and experimental interpretation.

Each normalization method relies on assumptions that impact interpretation, but caution is advised when considering the AP methods. AP1 and AP2 significantly distort the standard deviation of the data relative to GT data and AP3 alters the mean of the distribution relative to GT, as demonstrated by Gregor et al. [31].

There are many approaches that we did not consider here. In particular, similar evaluation is needed for approaches such as the evaluation of internal controls to compare data between multiple different experiments (or labs) and the comparison of non-exponential type gradients. In summary, when quantitative image data are used to infer mechanisms of morphogen patterning, the reported methods should include details about each image processing step, how the data were normalized, and a rationale for the choice of method. The increased attention to trade-offs in data normalization will improve our quantitative understanding of morphogen-mediated patterning and the cell biology of BMP signaling.

Acknowledgements

The Raftery laboratory is supported by NIH grant 5R01GM60501-13, and thanks the University of Nevada School of Medicine Department of Surgery Research Laboratories-Las Vegas, for the use of their LSCM. The Umulis laboratory wishes to acknowledge Purdue University for support.

References

[1] Araujo, H., Fontenele, M.R. and da Fonseca, R.N. (2011) Position matters: variability in the spatial pattern of BMP modulators generates functional diversity. Genesis 49, 698–718. 10.1002/dvg.20778.
[2] Umulis, D., O'Connor, M.B. and Blair, S.S. (2009) The extracellular regulation of bone morphogenetic protein signaling. Development 136, 3715–3728. 136/22/3715 [pii] 10.1242/dev.031534.
[3] Schwank, G. and Basler, K. (2010) Regulation of organ growth by morphogen gradients. Cold Spring Harb. Perspect. Biol. 2, a001669. 10.1101/cshperspect.a001669.
[4] Wartlick, O., Mumcu, P., Julicher, F. and Gonzalez-Gaitan, M. (2011) Understanding morphogenetic growth control – lessons from flies. Nat. Rev. Mol. Cell Biol. 12, 594–604. 10.1038/nrm3169 nrm3169 [pii].
[5] Harris, R.E. and Ashe, H.L. (2011) Cease and desist: modulating short-range Dpp signalling in the stem-cell niche. EMBO Rep. 12, 519–526. embor201180 [pii] 10.1038/embor.2011.80.
[6] Chen, S., Wang, S. and Xie, T. (2011) Restricting self-renewal signals within the stem cell niche: multiple levels of control. Curr. Opin. Genet. Dev. 21, 684–689. S0959-437X(11)00120-1 [pii] 10.1016/j.gde.2011.07.008.
[7] Ben-Zvi, D., Pyrowolakis, G., Barkai, N. and Shilo, B.Z. (2011) Expansion-repression mechanism for scaling the Dpp activation gradient in Drosophila wing imaginal discs. Curr. Biol. 21, 1391–1396. S0960-9822(11)00784-6 [pii] 10.1016/j.cub.2011.07.015.
[8] Bollenbach, T., Pantazis, P., Kicheva, A., Bokel, C., Gonzalez-Gaitan, M. and Julicher, F. (2008) Precision of the Dpp gradient. Development 135, 1137–1146. 135/6/1137 [pii] 10.1242/dev.012062.
[9] Hamaratoglu, F., de Lachapelle, A.M., Pyrowolakis, G., Bergmann, S. and Affolter, M. (2011) Dpp signaling activity requires Pentagone to scale with tissue size in the growing Drosophila wing imaginal disc. PLoS Biol. 9, e1001182. 10.1371/journal.pbio.1001182 PBIOLOGY-D-11-02056 [pii].
[10] Umulis, D.M., Shimmi, O., O'Connor, M.B. and Othmer, H.G. (2010) Organism-scale modeling of early Drosophila patterning via bone morphogenetic proteins. Dev. Cell 18, 260–274. S1534-5807(10)00010-9 [pii] 10.1016/j.devcel.2010.01.006.
[11] Wang, Y.C. and Ferguson, E.L. (2005) Spatial bistability of Dpp-receptor interactions during Drosophila dorsal-ventral patterning. Nature 434, 229–234. nature03318 [pii] 10.1038/nature03318.
[12] Wartlick, O., Mumcu, P., Kicheva, A., Bittig, T., Seum, C., Julicher, F. and Gonzalez-Gaitan, M. (2011) Dynamics of Dpp signaling and proliferation control. Science 331, 1154–1159. 331/6021/1154 [pii] 10.1126/science.1200037.
[13] Eldar, A., Dorfman, R., Weiss, D., Ashe, H., Shilo, B.Z. and Barkai, N. (2002) Robustness of the BMP morphogen gradient in Drosophila embryonic patterning. Nature 419, 304–308. 10.1038/nature01061 nature01061 [pii].
[14] Mizutani, C.M., Nie, Q., Wan, F.Y., Zhang, Y.T., Vilmos, P., Sousa-Neves, R., Bier, E., Marsh, J.L. and Lander, A.D. (2005) Formation of the BMP activity gradient in the Drosophila embryo. Dev. Cell 8, 915–924. S1534-5807(05)00140-1 [pii] 10.1016/j.devcel.2005.04.009.
[15] Kicheva, A., Pantazis, P., Bollenbach, T., Kalaidzidis, Y., Bittig, T., Julicher, F. and Gonzalez-Gaitan, M. (2007) Kinetics of morphogen gradient formation. Science 315, 521–525. 315/5811/521 [pii] 10.1126/science.1135774.
[16] Harris, R.E., Pargett, M., Sutcliffe, C., Umulis, D. and Ashe, H.L. (2011) Brat promotes stem cell differentiation via control of a bistable switch that restricts BMP signaling. Dev. Cell 20, 72–83. S1534-5807(10)00546-0 [pii] 10.1016/j.devcel.2010.11.019.
[17] Yakoby, N., Lembong, J., Schupbach, T. and Shvartsman, S.Y. (2008) Drosophila eggshell is patterned by sequential action of feedforward and feedback loops. Development 135, 343–351. dev.008920 [pii] 10.1242/dev.008920.
[18] Aegerter-Wilmsen, T., Aegerter, C.M., Hafen, E. and Basler, K. (2007) Model for the regulation of size in the wing imaginal disc of Drosophila. Mech. Dev. 124, 318–326. S0925-4773(06)00220-6 [pii] 10.1016/j.mod.2006.12.005.
[19] Peng, H. (2008) Bioimage informatics: a new area of engineering biology. Bioinformatics 24, 1827–1836. 10.1093/bioinformatics/btn346.
[20] Akiyama, T., Kamimura, K., Firkus, C., Takeo, S., Shimmi, O. and Nakato, H. (2008) Dally regulates Dpp morphogen gradient formation by stabilizing Dpp on the cell surface. Dev. Biol. 313, 408–419. S0012-1606(07)01485-6 [pii] 10.1016/j.ydbio.2007.10.035.

[21] Ogiso, Y., Tsuneizumi, K., Masuda, N., Sato, M. and Tabata, T. (2011) Robustness of the Dpp morphogen activity gradient depends on negative feedback regulation by the inhibitory Smad, Dad. Dev. Growth Differ. 53, 668–678. 10.1111/j.1440-169X.2011.01274.x.
[22] Gregor, T., Wieschaus, E.F., McGregor, A.P., Bialek, W. and Tank, D.W. (2007) Stability and nuclear dynamics of the bicoid morphogen gradient. Cell 130, 141–152. S0092-8674(07)00663-0 [pii] 10.1016/j.cell.2007.05.026.
[23] He, F., Wen, Y., Deng, J., Lin, X., Lu, L.J., Jiao, R. and Ma, J. (2008) Probing intrinsic properties of a robust morphogen gradient in Drosophila. Dev. Cell 15, 558–567. S1534-5807(08)00377-8 [pii] 10.1016/j.devcel.2008.09.004.
[24] Hengenius, J.B., Gribskov, M., Rundell, A.E., Fowlkes, C.C. and Umulis, D.M. (2011) Analysis of gap gene regulation in a 3D organism-scale model of the Drosophila melanogaster embryo. PLoS ONE 6, e26797. 10.1371/journal.pone.0026797 PONE-D-11-07665 [pii].
[25] Jaeger, J., Surkova, S., Blagov, M., Janssens, H., Kosman, D., Kozlov, K.N., Manu, Myasnikova.E., Vanario-Alonso, C.E., Samsonova, M., et al. (2004) Dynamic control of positional information in the early Drosophila embryo. Nature 430, 368–371. 10.1038/nature02678 nature02678 [pii].
[26] Houchmandzadeh, B., Wieschaus, E. and Leibler, S. (2002) Establishment of developmental precision and proportions in the early Drosophila embryo. Nature 415, 798–802. 10.1038/415798a 415798a [pii].
[27] Kanodia, J.S., Rikhy, R., Kim, Y., Lund, V.K., DeLotto, R., Lippincott-Schwartz, J. and Shvartsman, S.Y. (2009) Dynamics of the Dorsal morphogen gradient. Proc. Natl. Acad. Sci. U S A 106, 21707–21712. 0912395106 [pii] 10.1073/pnas.0912395106.
[28] Reeves, G.T. and Stathopoulos, A. (2009) Graded dorsal and differential gene regulation in the Drosophila embryo. Cold Spring Harb. Perspect. Biol. 1, a000836. 10.1101/cshperspect.a000836.
[29] Cumming, G., Fidler, F. and Vaux, D.L. (2007) Error bars in experimental biology. J. Cell Biol. 177, 7–11. jcb.200611141 [pii] 10.1083/jcb.200611141.
[30] He, F., Wen, Y., Cheung, D., Deng, J., Lu, L.J., Jiao, R. and Ma, J. (2010) Distance measurements via the morphogen gradient of Bicoid in Drosophila embryos. BMC Dev. Biol. 10, 80. 1471-213X-10-80 [pii] 10.1186/1471-213X-10-80.
[31] Gregor, T., Tank, D.W., Wieschaus, E.F. and Bialek, W. (2007) Probing the limits to positional information. Cell 130, 153–164. S0092-8674(07)00662-9 [pii] 10.1016/j.cell.2007.05.025.
[32] Crauk, O. and Dostatni, N. (2005) Bicoid determines sharp and precise target gene expression in the Drosophila embryo. Curr. Biol. 15, 1888–1898. S0960-9822(05)01140-1 [pii] 10.1016/j.cub.2005.09.046.
[33] Allchin, D. (2001) Error Types. Perspect. Sci. 9, 21.
[34] Ashburner, M., Golic, K.G. and Hawley, R.S. (2005) Drosophila: a laboratory handbook, 2nd edn, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y..
[35] Cohen, S. (1993) Imaginal disk development in: The Development of Drosophila melanogaster (Bate, M. and Martine-Arias, A., Eds.), pp. 747–841, Cold Spring Harbor Laboratory Press.
[36] Parker, J. (2011) Morphogens, nutrients, and the basis of organ scaling. Evol. Dev. 13, 304–314. 10.1111/j.1525-142X.2011.00481.x.
[37] Raftery, L.A. and Umulis, D.M. (2012) Regulation of BMP activity and range in Drosophila wing development. Curr. Opin. Cell Biol. 24, 158–165.
[38] Kahlem, P. and Newfeld, S.J. (2009) Informatics approaches to understanding TGFbeta pathway regulation. Development 136, 3729–3740. 136/22/3729 [pii] 10.1242/dev.030320.
[39] Affolter, M. and Basler, K. (2007) The Decapentaplegic morphogen gradient: from pattern formation to growth regulation. Nat. Rev. Genet. 8, 663–674. nrg2166 [pii] 10.1038/nrg2166.
[40] Dutko, J.A. and Mullins, M.C. (2011) SnapShot: BMP signaling in development. Cell 145 (636), e631–632. S0092-8674(11)00487-9 [pii] 10.1016/j.cell.2011.05.001.
[41] Waters, J.C. (2009) Accuracy and precision in quantitative fluorescence microscopy. J. Cell Biol. 185, 1135–1148. jcb.200903097 [pii] 10.1083/jcb.200903097.
[42] Conchello, J.A. and Lichtman, J.W. (2005) Optical sectioning microscopy. Nat. Methods 2, 920–931. nmeth815 [pii] 10.1038/nmeth815.
[43] Lichtman, J.W. and Conchello, J.A. (2005) Fluorescence microscopy. Nat. Methods 2, 910–919. nmeth817 [pii] 10.1038/nmeth817.
[44] North, A.J. (2006) Seeing is believing? A beginners' guide to practical pitfalls in image acquisition. J. Cell Biol. 172, 9–18. jcb.200507103 [pii] 10.1083/jcb.200507103.
[45] Kicheva, A., Holtzer, L., Wartlick, O., Schmidt, T. and Gonzalez-Gaitan, M. (2011) Quantitative imaging of morphogen gradients in Drosophila imaginal discs in: Imaging in Developmental Biology: A Laboratory Manual (Sharpe and Wang, Eds.), pp. 550–553, Cold Spring Harbor Press.
[46] Morrison, A., Scheeler, M., Dubuis, J. and Gregor, T. (2011) Quantifying the Bicoid morphogen gradient in living fly embryos in: Imaging in Developmental Biology (Sharpe and Wang, Eds.), pp. 523–532, Cold Spring Harbor Press.
[47] Brand, A.H. and Perrimon, N. (1993) Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. Development 118, 401–415.
[48] Cao, J., Pellock, B., White, K. and Raftery, L. (2006) A commercial phospho-Smad antibody that detects endogenous BMP signaling in *Drosophila* tissues. Dros. Inform. Serv. 89, 131–135.
[49] Peluso, C.E., Umulis, D., Kim, Y.J., O'Connor, M.B. and Serpe, M. (2011) Shaping BMP Morphogen Gradients through Enzyme-Substrate Interactions. Dev. Cell 21, 375–383. 10.1016/j.devcel.2011.06.025.
[50] Tanimoto, H., Itoh, S., ten Dijke, P. and Tabata, T. (2000) Hedgehog creates a gradient of DPP activity in Drosophila wing imaginal discs. Mol. Cell 5, 59–71. S1097-2765(00)80403-7 [pii].
[51] Bangi, E. and Wharton, K. (2006) Dpp and Gbb exhibit different effective ranges in the establishment of the BMP activity gradient critical for Drosophila wing patterning. Dev. Biol. 295, 178–193. S0012-1606(06)00210-7 [pii] 10.1016/j.ydbio.2006.03.021.
[52] Myasnikova, E., Surkova, S., Panok, L., Samsonova, M. and Reinitz, J. (2009) Estimation of errors introduced by confocal imaging into the data on segmentation gene expression in Drosophila. Bioinformatics 25, 346–352. btn620 [pii] 10.1093/bioinformatics/btn620.
[53] Vuilleumier, R., Springhorn, A., Patterson, L., Koidl, S., Hammerschmidt, M., Affolter, M. and Pyrowolakis, G. (2010) Control of Dpp morphogen signalling by a secreted feedback regulator. Nat. Cell Biol. 12, 611–617. ncb2064 [pii] 10.1038/ncb2064.
[54] Baena-Lopez, L.A., Nojima, H. and Vincent, J.-P. (2012) Integration of morphogen signaling within the growth regulatory network. Curr. Opin. Cell Biol. 24, 166–172, http://dx.doi.org/10.1016/j.ceb.2011.12.010.